STAT 300 Chapter 8 Hypothesis Tests

Zepu Zhang
April 5, 2012

Sections 8.1 and 8.4 are more general; section 8.2 is particular cases. A good understanding of sections 8.1 and 8.4 will make 8.2 "piece of cake". Skip 8.3 and 8.5.

1 Hypotheses and test procedures

1.1 Hypotheses

Hypothesis (in this chapter) is a statement about an unknown distribution parameter θ, for example, $\theta = 1$; $5.2 < \theta < 5.8$; $\theta > 0$.

A "null" hypothesis, H_0, is a statement that is taken by default unless sample data provide strong evidence against it. An "alternative" hypothesis, H_a, is the statement that the data suggest other than the null.

Hypothesis tests look into sample data and see whether there is strong evidence that H_0 should be rejected.

There are two possible conclusions of a test:
(1) reject H_0 (and take H_a);
(2) fail to reject H_0 (or "accept" H_0).

Sometimes for convenience we say "accept H_0" to mean "fail to reject H_0", but note the following interpretation: failing to reject H_0 is a lack of disproof, which is not the same as proof. We do not say "we believe H_0 is true". We're not sure about that; it's just that evidence against it is not strong enough (by a prescribed criterion).

In this chapter, we always take H_0 to be an equality to a particular number, say θ_0:
H_0: $\theta = \theta_0$

H_a has three possible forms:
H_a: $\theta > \theta_0$
H_a: $\theta < \theta_0$

H_a: $\theta \neq \theta_0$

We see that H_0 and H_a are not always "complementary" to each other. For example, $\theta \leq \theta_0$ and $\theta > \theta_0$ would complement each other but $\theta = \theta_0$ and $\theta > \theta_0$ do not. In fact, when we test $H_0 : \theta = \theta_0$ vs $H_a : \theta > \theta_0$, we are really choosing between $\theta \leq \theta_0$ and $\theta > \theta_0$, but we use $H_0 : \theta = \theta_0$ for technical reasons. Some explanations appear in section 8.1, but don't worry about it.

1.2 Logic of the test procedure

Take a function of the sample data, say T, called "test statistic". The test statistic is a random variable, because the data are a random sample.

The test statistic is a deliberate choice such that it is connected with θ and H_0. In particular, the distribution of T is known given that H_0 is assumed true. This distribution is called the null distribution of T.

Example To test about μ, what about taking $T = \overline{X}$? Do we know the distribution of this T when $\theta = \theta_0$? (Suppose we know the population distribution is $N(\mu, \sigma^2)$ where σ^2 is known.)

We identify a rejection region, i.e. a set of values of T; let's denote it by R. (If T is continuous, R is an interval or the union of more than one interval; if T is discrete, R is a set of particular values.) If the observed value of T (obtained from the actually observed data) lies in R, we reject H_0.

This is called our decision rule: we reject H_0 if $T \in R$, and accept H_0 if $T \notin R$.

We always choose R to consist of the most extreme (i.e. unlikely) values in the null distribution of T. For example, if the null distribution is normal, we may take R to be the union of the 2.5% left tail and 2.5% right tail.

When the observed value of T falls in R, there is a contradiction—if H_0 were true, then we know the null distribution of T, and R consists of extreme values in this distribution; we shouldn't have observed values in R (because they're extreme, i.e. very unlikely to happen), but the observed value is in R! Something is suspicious! What caused this awkward situation? The probability theory is sound; the only suspect is the assumption that H_0 is true. Hence we conclude H_0 is not true.

1.3 Errors in tests

Definition Type I error: wrongly reject H_0, while in fact H_0 is true. Type II error: wrongly accept H_0, while in fact H_0 is false.

Source of error: Because we check a data sample, not the whole population, there is always a chance that the sample is "unrepresentative" of the population and "fools" us to make a wrong judgement about the population (or, the unknown population parameter, θ).

Suppose in fact H_0 is true, then a "representative" sample will result in a T value that is unlikely to belong in R. However, an unusual data sample could result in a T in R, leading us to mistakenly reject H_0—error! This is called a type I error.

Suppose in fact H_0 is false, then the distribution of T is not the null distribution. In the true distribution of T, the rejection region R is not necessarily an extreme region, and $\overline{R}$ (the acceptance region) is not necessarily an extreme region, either. Hence, the observed T value may well fall in R, or $\overline{R}$. If it happens to fall in $\overline{R}$, we will mistakenly accept H_0—error! This is called a type II error.

Note When our conclusion is rejection, we could have made a type-I error; type-II error is meaningless in this situation. When our conclusion is acceptance, we could have made a type-II error; type-I error is meaningless in this situation.

1.4 Probability of errors

If in fact H_0 is true, i.e. $\theta = \theta_0$, then we know the distribution of T, hence we know the probabilities $P(T \in R)$ and $P(T \notin R)$. The probability of rejecting H_0, hence making a type I error, is

$$\alpha = P(T \in R \,|\, H_0), \quad \text{i.e. } \alpha = P(T \in R \,|\, \theta = \theta_0)$$

Because θ_0 (the hypothesized value) is given, α can be calculated.

If in fact H_0 is false, the probability of accepting H_0, hence making a type II error, is

$$\beta = P(T \in \overline{R} \,|\, H_a)$$

To calculate this probability, we need to know the distribution of T under H_a. This requires knowing the exact value of θ. If H_a is in the forms like $\theta > \theta_0$, $\theta < \theta_0$, or $\theta \neq \theta_0$, it's not enough; we don't know the exact value of θ, hence we can't calculate β. (In a test with $H_0 : \theta = a$ vs $H_a : \theta = b$, we would be able to calculate β. We don't do that kind of tests in this course.)

In general, R is chosen to be the most extreme part of the null distribution that has probability α. (Being "extreme" here means unlikely under H_0 but likely under H_a.) Typical levels of α are 0.1, 0.05, 0.01. α is called Significance level. Note that we are able to pick R according to any specified value of α because we know the null distribution.

Relations between α and β:

1. α is in our control; we usually set (the upper limit of) α to a desired level (depending on how much type I error is acceptable to us), determine R based on the null distribution and the required α.

2. β is unknown and can't be calculated, because the actual value of θ is unknown.

3. α and β vary in opposite directions: setting α to a smaller value will lead to larger β.

2 The critical value approach to tests about a population mean

The procedure is pretty routine. Always do your problems in this procedure and clearly label the steps as such:

1. State the hypotheses: H_0 and H_a, and the specified α level.

2. Define (or "choose") a test statistics T; state the null distribution of T.

   (T is a statistic of the data sample. Write it in terms of general symbols like $X_1, \dots, X_n$ as well as unknown population parameters; do not yet plug in actual data values.)

3. Find the rejection region R. This may be stated in either of the following ways:

   (a) State the "rejection region" as an interval, e.g. $(-\infty, -1.65)$, or $(-\infty, -1.96) \cup (1.96, \infty)$.

   (b) State the "decision rule", e.g., "reject H_0 if $|T| > 1.96$ and accept H_0 otherwise".

4. Calculate the value of T using the sample data. Call the value T^*.

5. State your conclusion according to the rejection region (or decision rule) and the value T^*. Conclusion is either

   - "Fail to reject H_0" (or "accept H_0"), or
   - "Reject H_0".

   It is desirable to re-state the conclusion in the terminology and context of the actual problem.

Steps 1–3 are done without using the actual data (and can be done even before data are obtained).

Note The rejection region (or the decision rule) is determined based on (1) the null distribution of T; (2) the alternative hypothesis H_a; (3) the specified significance level, α. Specifically, the rejection region is the set of the "most extreme values of T in its null distribution" such that these extreme values take up probability α and they are "extreme in the direction that suggests H_a".

For example, suppose $H_0 : \mu = 3$, $H_a : \mu > 3$, and $T = \overline{X}$. Then an extremely large $\overline{X}$ "suggests" H_a. An extremely small $\overline{X}$, although throws doubt on H_0, does not provide evidence for H_a. Therefore the rejection region for this problem consists of extremely large $\overline{X}$ values.

To test about a population mean μ, our test statistic is based on the sample mean $\overline{X}$. In practice we do not take

$\overline{X}$ directly as the test statistic; rather we take a standardized (or stuentized) version of it, whose distribution is simpler.

2.1 Case 1: known variance, normal population (z test)

In this case, we know $\overline{X}$ is normal,

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

hence

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1) \tag{1}$$

In the test we assume the null hypothesis $H_0 : \mu = \mu_0$ is true, therefore the μ in the formula above is replaced by μ_0. We take the test statistic

$$T = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$$

Its null distribution is $N(0, 1)$.

Note that T is a measure of the distance of $\overline{X}$ from the hypothesized population mean μ_0 in multiples of $\sigma_{\overline{X}}$ (i.e. $\sigma/\sqrt{n}$). An extremely large T results from a $\overline{X}$ that is much larger than μ_0, suggesting the true μ is actually some value above μ_0. An extremely small T results from a $\overline{X}$ that is much smaller than μ_0, suggesting the true μ is actually some value below μ_0.

Let the value of T calculated based on the available data set be denoted by T^*. If T^* falls in the rejection region R, we reject H_0; otherwise, we do not reject H_0.

Decision rules:
$H_a : \mu < \mu_0$: reject H_0 if $T^* < -z_\alpha$
$H_a : \mu > \mu_0$: reject H_0 if $T^* > z_\alpha$
$H_a : \mu = \mu_0$: reject H_0 if $|T^*| > z_{\alpha/2}$

The values z_α, $z_{\alpha/2}$ are called the "critical values".

Example Ex. 8.6 in Chap. 8.2.

2.2 Case 2: known variance, large sample (z test)

If we don't know whether the population distribution is normal, but the sample size is large (say $n \geq 30$), then by the Central Limit Theorem, the same test statistic again has a standard normal null distribution. Therefore the procedure and decision rules are the same as in the z test.

2.3 Case 3: unknown variance, normal population (t test)

In this case we take the test statistic

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

and we know its null distribution is t_{n-1}.

The test is as before, except that the critical values are t values instead of z values.

Decision rules:
$H_a : \mu < \mu_0$: reject H_0 if $T^* < -t_{\alpha,n-1}$
$H_a : \mu > \mu_0$: reject H_0 if $T^* > t_{\alpha,n-1}$
$H_a : \mu = \mu_0$: reject H_0 if $|T^*| > t_{\alpha/2,n-1}$

Example Ex. 8.9 in Chap. 8.2.

2.4 Case 4: unknown variance, large sample

In this case we use the t test.

3 P-values

Shortcomings of the "critical value" approach:

First, α (the acceptable type-I error probability) is specified. Then, rejection region is determined based on α. Finally, a conclusion is made based on the actual value of the test statistic and its relation with the rejection region. In other words, this procedure sets a criterion for the "extremeness" of the test statistic T, then proceeds to report whether the observed T^* is extreme (hence provides strong evidence against H_0) according to this criterion.

Problem: what if someone else prefers a different α (i.e. has a different opinion on how high a risk of type-I error is acceptable)? Re-do the whole test?

α is a pre-specified criterion about what values of T are considered "extreme" (or unlikely). When we report a test conclusion based on this pre-specified α, the conclusion does not reveal how close or far the actually observed T^* is from the borderline of the criterion. For example, if the conclusion is "rejection", we only know T^* is extreme according to the pre-specified α, but we don't know whether it just barely satisfies this criterion, or it is very extreme such that had we specified a much smaller α, the T^* would still be extreme according to that smaller α.

Alternative:

We report an "extremeness" measure of the observed T^* but do not conclude whether this is "extreme enough" to overthrow H_0; the judgement is left to the researcher who is free to use whatever level of α. With the same P-value, different people may come to different conclusions depending on how high a probability of type-I error they are willing to accept.

This "measure of extremeness" is called P-value. It is the probability, given H_0 is true, of observing a T value that is at least as extreme as or more extreme than the actually observed value. Here "extreme" means in the direction that suggests H_a is true.

Another way to define the P-value: it is the smallest α level on which H_0 would be rejected.

Making a conclusion by comparing α and P: reject H_0 if $p < \alpha$.

Remember a smaller P-value is stronger evidence against H_0 (and for H_a).

3.1 P-values for z tests

In Chap. 8.4.

Example Ex. 8.17 in Chap. 8.4.

3.2 P-values for t tests

In Chap. 8.4.

Example Ex. 8.18 in Chap. 8.4.